<Vorname Nachname [et. al.]>(Hrsg.): < Buchtitel>,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn <Jahr>

Knowledge Discovery in Big Data: Herausforderungen durch Big Data im Prozess der Wissensgewinnung am Beispiel des CRISP-DM

Thomas Göpfert[1], Andreas Breiter[2]

Abstract: Der Prozess valide, neuartige, potenziell nutzbare und verständliche Muster in Daten zu finden, wird als Knowledge Discovery in Database Prozess bezeichnet (KDD-Prozess). Die diesem Prozess zu Grunde liegende Datenbasis unterliegt einem ständigen Wandel. Doug Laney erkannte die Eigenschaften Volume, Variety und Velocity als neue Herausforderungen für IT-Organisationen. Heute werden diese Herausforderungen unter dem Begriff Big Data zusammengefasst. Die Auswirkungen von Big Data auf den KDD-Prozess sind bisher unzureichend untersucht. Ziel dieser Arbeit war es, die Herausforderungen durch Big Data am Beispiel des CRISP-DM, eines der am meisten genutzten KDD-Prozessmodelle, zu analysieren. Durch ein systematisches Literaturreview wurden elementare Herausforderungen identifiziert und den Prozessschritten des Prozessmodells zugeordnet. Die Ergebnisse konnten mittels Experteninterviews verifiziert werden. Neben der Identifikation zentraler Herausforderungen wurde deutlich, dass CRISP-DM bei der Analyse von Big Data Gültigkeit hat, aber zentrale Herausforderungen, vor allen in den Phasen der Datenvorverarbeitung, beachtet werden müssen.

Keywords: Big Data, Knowledge Discovery in Database, KDD, KDDM, CRSIP-DM, Data Mining

1 Einleitung

Aktuelle Studien gehen davon aus, dass die vorhanden Datenmenge (die Summe aller gespeicherten Daten) 2013 4,4 Zettabyte (1 Zettabyte sind 1021 Bytes) betrug. Es wird geschätzt, dass sich diese Datenmenge alle zwei Jahre verdoppelt. 2020 wird mit einem Datenvolumen von 44 Zettabyte gerechnet. IDC nennt diese Datenmenge in ihrer gleichnamigen Studie das „Digitale Universum“ und geht davon aus, dass diese Daten wie Dinge im physikalischen Universum auch, in ganz unterschiedlicher Form vorliegen [ID14].

Dieses Datenvolumen (Volume), die Datenvielfältigkeit (Variety) und die Geschwindigkeit in welcher Daten (Velocity) erzeugt werden, erkannte Doug Laney als neue Herausforderungen für IT-Organisationen [La01]. Heute bildet diese Beschreibung eine der geläufigsten Definitionen für Big Data [Ki14]. Die oben erwähnte Studie des

[1] Universität Bremen, AG Informationsmanagement, Bibliotheksstraße 1 , 28359 Bremen, thg@informtik.uni-bremen.de
[2] Universität Bremen, Institut für Informationsmanagement Bremen GmbH, Am Fallturm 1, 28359 Bremen , abreiter@ifib.de

IDC geht davon aus, dass 22 Prozent der Daten 2013 und 37 Prozent der Daten 2020 einen Nutzen bringen, wenn sie mit Schlagwörtern versehen und analysiert werden [ID14]. „Die intelligente Auswertung kann Organisationen wichtige Daten liefern. […] Es ist offensichtlich, dass solche Unternehmen einen Wettbewerbsvorteil erlangen, die aus der Vielzahl der Daten geschäftsrelevante Informationen filtern können“ [BI14]. Chen und Zhang beschreiben typische Anwendungen von Big Data in den Feldern Commerce and Business (Wirtschaft), Scientific Research (Wissenschaft) und Society Administration (Öffentliche Verwaltung) [PZ14]. Dabei macht nicht die Analyse der Daten an sich, sondern das Verstehen der Daten (d.h. u.a. das Anreichern der Daten mit Kontextinformationen), die Sicherstellung der Datenqualität und die Vorverarbeitung der Daten (zur Vorbereitung für die eigentliche Analyse) einen Großteil des Aufwandes bei der Datenanalyse aus. So nimmt Franks an, dass 95 bis 100 Prozent des Aufwands der gesamten Datenanalysen alleine darin besteht, diese Arbeitsschritte zu erledigen [Fr12].

Fayyad u.a. analysieren lange vor der intensiven Diskussion des Begriffes Big Data, dass Data Mining (also die eigentliche Anwendung von Algorithmen) nur ein Teil eines Prozesses ist, den sie Knowledge Discovery in Database (KDD-Prozess) nennen [FPS96]. Auch große Unternehmen haben dies erkannt und einen standardisierten Prozess zur Datenanalyse entwickelt [IB10]. Dieser Cross-Industry Standard Process for Data Mining (CRISP-DM) ist das heute am weitesten verbreitete Prozessmodell zur Datenanalyse [Be10].

Es ist anzunehmen, dass diese Prozessmodelle nicht die angesprochen Entwicklung der Herausforderungen von Volume, Variety und Velocity berücksichtigen. So entstanden diese Prozesse der Wissensgewinnung bereits in den Jahren 1996 (KDD-Prozess) und 2000 (CRISP-DM), während z.B. die meistverwendete Beschreibung der Eigenschaften von Big Data von Doug Laney auf das Jahr 2001 zurückgeht. Auf Grundlage dieser Annahme hat dieser Beitrag das Ziel, die Herausforderungen durch Big Data in den jeweiligen Phasen des Prozess der Wissensgewinnung am Beispiel des CRSIP-DM systematisch zu identifizieren und zu benennen.

2 Cross-Industry Standard Process for Data Mining

Der Cross-Industry Standard Process for Data Mining (CRISP-DM) ist ein industrie-, werkzeug- und anwendungsneutrales Prozessmodell. Es ist laut Berthold u.a. das am Weitesten verbreitete und eingesetzte Modell [Be10]. Der CRISP-DM 1.0 Step-by-step data mining guide beschreibt vier in ihrem Abstraktionslevel unterschiedliche Ebenen. Die obersten zwei Ebenen werden als CRISP-DM reference model bezeichnet, die zwei unteren Ebenen als CRISP-DM user guide. Jede Phase besteht aus generischen Aufgaben, welche zusammen die Ebene darunter bilden. Die Phasen mit ihren jeweiligen generischen Aufgaben bilden das CRIPS-DM reference model. Die Abstraktion dieser Phasen und Aufgaben ist so gewählt, dass diese für alle Data Mining Anwendungen Gültigkeit haben (Vollständigkeit des Modells) und zukünftige Entwicklungen, z.B. neue

Modellierungstechniken, diese Phasen und Aufgaben nicht beeinflussen (Stabilität des Modells) [IB10].

Die Phasen des CRISP-DM reference model sind Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment. Ziel der Phase Business Understanding ist es, ein Verständnis über das Projekt aufzubauen. Dazu zählen die Anforderungen aus Sicht des Kunden, die gegebenen Faktoren, welche zu berücksichtigen sind die Anforderungen für Data Mining Aufgaben, welche sich aus den Zielen des Kunden ableiten und ein erster Projektplan. Ein Verständnis der Daten zu erlangen, ist das globale Ziel der Phase Data Understanding. Hierfür werden die Daten gesammelt, beschrieben, erste Einblicke gewonnen, interessante Daten identifiziert, ggf. erste Hypothesen aufgestellt und die Qualität der Daten überprüft. Am Ende der Phase Data Preparation stehen die Datensätze, mit welchen in der Phase Modeling (sie sind Eingabe für die Data Mining Algorithmen) gearbeitet wird. In der Phase Modeling werden verschiedene Modellierungstechniken gewählt und angewandt. Die Ergebnisse der Modeling Phase werden mit den Zielen des Kunden in der der Phase Evaluation verglichen. Ausgehend von der Annahme, dass die Generierung und die Evaluation eines Modells nicht das Projekt abschließen, sondern vielmehr die Ergebnisse zum Zwecke der Wissensgewinnung aus den Daten genutzt werden sollen, muss das Modell Anwendung finden, auch als Auslieferung (Deployment) bezeichnet. Hierzu müssen Ergebnisse ggf. in einer nutzbaren Form organisiert und dargestellt werden [IB10].

3 Stand der Forschung

Datensätze werden als Big Data bezeichnet, wenn es mit aktueller Technologie schwierig ist, diese zu erheben, zu pflegen, zu analysieren und zu visualisieren [PZ14], [Ja11]. Eine der geläufigsten Beschreibung der Eigenschaften von Big Data lieferte Doug Laney in seiner Arbeit „3D Data Management: Controlling Data Volume, Velocity, and Variety“. Er erkannte neue Herausforderungen für IT-Organisationen im Umgang mit Daten in den Dimensionen Datenvolumen (Terabyte bis Petabyte an vorhandenen Daten), der Vielfältigkeit (in ihrer Art, d.h. strukturierte, unstrukturierte und semi-strukturierte Daten) der Daten und der Geschwindigkeit in der Daten erzeugt und analysiert werden (nahezu in Echtzeit) [La01]. Aktuelle empirische Studien wie [Gl12] oder [BI12] zeigen Herausforderungen bzw. Hürden bei der Einführung bzw. Umsetzung von Big Data Projekten in der Praxis. Dabei ist eine Mischung aus organisatorischen und technischen Faktoren festzustellen [Gl14].

Einige Beiträge untersuchen Big Data anhand oder mittels einzelner Prozessschritte, wie sie im KDD-Prozess oder im CRISP-DM vorkommen. So analysier Rajpurohit, dass ein Hand-in-Hand zwischen Technologie und Business Understanding nötig und dazu ein Verständnis des KDD-Prozess Grundlage ist, um durch Big Data zu profitieren. Er führt dabei das Prozessmodell von Fayyad u.a. an und erklärt die einzelnen Schritte ohne dabei Modifikationen oder Herausforderungen bzgl. Big Data hervorzuheben. Allgemein

stellt er allerdings deutlich die Schwierigkeit heraus, dass die Deckung der Kosten für die Einführung einer Big Data Lösung (Return of Investment) schwer zu kalkulieren sind [Ra13].

Auch Begoli und Horey stellen heraus, dass durch Big Data die Datenanalyse mittels KDD-Prozess die konsequente Lösung ist. Sie heben hervor, dass ein effektiver Einsatz dieser Methode allerdings technologische und organisatorische Erfahrung voraussetzt. Aus ihren praktischen Erfahrungen beim Oak Ridge National Laboratory (ORNL) heraus identifizieren sie drei wesentliche Design Prinzipien, deren Anwendung aus ihrer Sicht wichtig ist, um ökonomische, umfassende, flexible und sichere Lösungen für die Big Data Anforderungen im öffentlichen Sektor zu generieren. Diese Prinzipien sind: die Unterstützung einer Vielzahl von Analysemethoden, das Bereitstellen von Lösungen zur Speicherung und Verarbeitung von Daten in allen Prozessschritten und sie Verfügbarkeit und Verständlichkeit resultierenden Daten [BH12].

Eine ausführliche Analyse von Big Data präsentieren Chen und Zhang in ihrer Arbeit „Data-intensive applications, challenges, techniques and technologies: A survey on Big Data“. Ein Teil dieser Arbeit besteht darin, die Herausforderungen durch Big Data anhand der Prozessschritte Datenerhebung und -speicherung, Datenübermittlung, Datenpflege, Datenanalyse und Datenvisualisierung zu beschreiben. Dabei sehen die Autoren in den Schritten Datenerhebung und Datenspeicherung und die Gewährleistung des Zugangs zu Big Data als eine Aufgabe mit top Priorität im Knowledge Discovery Prozess. Die Daten sollten einfach und sofort zugänglich sein. Die Vorverarbeitung (Datenbereinigung, Datentransformation, Datenkatalogisierung) der Daten vor der Analyse ist notwendig um Data Mining Methoden anzuwenden. Auch für die eigentliche Datenanalyse müssen neue Methoden entwickelt werden. Nicht nur für die Speicherung, Pflege und Analyse, sondern auch für die Visualisierung werden laut Chen und Zhang neue Methoden benötigt [PZ14].

4 Methode

Basierend auf der Beobachtung, dass nach bestem Gewissen keine Analyse der Herausforderungen durch Big Data im CRISP-DM existiert, dass aber durchaus bereits zahlreiche Herausforderungen durch Big Data benannt wurden, resultiert folgendes Vorgehen.

Zur Identifikation der bereits benannten Herausforderungen wurde ein systematisches Literaturreview, angelehnt an [Ki04], durchgeführt. Anhand der Forschungsfrage wurde eine Reviewstrategie abgeleitet und dokumentiert. Diese basiert auf den Phasen des CRISP-DM und wurden in sieben verschiedene Suchen (je Phasen plus eine allgemeine Suche nach Herausforderungen) mit unterschiedlichen Suchstrings aufgeteilt. Das Review wurde mittels der Suchmaschine Web of Science der Firma Thomson Reuters[3]

[3] Webseite der Suchmaschine Web of Science: webofknowledge.com

durchgeführt. Dabei wurden die Suchergebnisse auf das Erscheinungsjahr 2012-2014 und den zugeordneten Wissenschaftsbereich „Computer Science“ beschränkt. Die Suchergebnisse wurden nach Relevanz und Anzahl der Zitationen sortiert und eine entsprechende Anzahl an Treffern je Suche zur näheren Analyse ausgewählt. Es konnte so 28 Arbeiten zur näheren Analyse herangezogen werden. Mittels zusammenfassender Inhaltsanalyse nach [Ma10] wurden Herausforderungen identifiziert, abstrahiert, generalisiert und den Phasen des CRISP-DM zugeordnet. Eine Zuordnung wurde vorgenommen falls die Herausforderung einer der Aufgaben der jeweiligen Phase tangiert. Darüber hinaus wurden übergreifende Herausforderungen gebündelt. Das daraus entstandene Kategoriensystem setzt sich dementsprechend aus deduktiven und einer induktiv gewonnenen Kategorie zusammen.

Aufbauend auf den Erkenntnissen aus dem systematischen Literaturreview wurden zwei leitfadengestützte Experteninterviews durchgeführt. Diese wurden transkribiert und ebenfalls anhand der zusammenfassenden Inhaltsanalyse und mittels des im systematischen Literaturreview entstandenen Kategoriensystems analysiert. Für die Interviews konnten ein Experte aus dem Bereich Wissenschaft und ein Experte aus dem Bereich Wirtschaft, also zwei der drei durch Chen und Zhang genannten Anwendungsbereiche von Big Data [PZ14] gewonnen werden. Im Bereich Wirtschaft konnte das Interview mit einem Projektleiter eines mittelständischen Unternehmens, spezialisiert auf die Beratung und Durchführung von Projekten im Bereich Business Intelligence, Data Warehouse und Berichtswesen durchgeführt werden. Im Bereich Wissenschaft wurde ein Sozialwissenschaftler mit dem Fokus auf Analyse von Twitterdaten interviewt. Beide Interviewpartner führen oder führten bereits Big Data Projekte durch, die der genannten Definition von Big Data entsprechen.

5 Ergebnisse

Insgesamt wurden 208 elementare Herausforderungen mittels systematischem Literaturreview und Experteninterviews identifiziert und den Kategorien, also den Phase des CRSIP-DM reference model, zugeordnet. Im Folgenden werden die zentralen Herausforderungen durch Big Data der einzelnen Phasen des CRISP-DM dargestellt. Eine jeweils einleitende Tabelle fasst wesentliche Herausforderungen zusammen und gibt an ob diese durch das systematische Literaturreview (L) oder durch die Experteninterviews (I) identifiziert wurden und/oder dem Stand der Forschung (F) entsprechen.

5.1 Business Understanding

Herausforderung	**L**	**I**	**F**
Business Case, Business Ziele und abgeleitete Anforderungen definieren	Ja	Ja	Ja
Zusammenarbeit Management und Kunden	Ja	Ja	Ja
Datenschutz und Ethik beachten	Ja, nicht Ethik	Ja	Ja, nicht Ethik
Projektplanung und Kostenschätzung/-planung	Ja	Ja, Kosten nein	Ja

Tab. 1: Zentrale Herausforderungen der Phase Business Understanding

Eine zentrale Aufgabe der Phase Business Understanding ist die Festlegung der Business Ziele (Aufgabe Determine Business Objectives). Eine erste Herausforderung dieser Phase stellt die Definition von Geschäftsszenarios dar, in welchen durch Big Data gewonnene nützliche Informationen eine Rolle spielen und einen Vorteil bringen. Ziele eines Big Data Projektes müssen klar definiert und insbesondere von allen Beteiligten verstanden werden. Diese Ziele sollten u.a. auch die Möglichkeiten aktueller Technologien berücksichtigen. Dazu muss auch der Kunde aktuelle Technologien verstanden haben. Big Data Projekte haben meist einen großen, ggf. experimentellen Charakter und bedürfen der Unterstützung des Managements. Es hat sich in den Interviews gezeigt, dass eng mit der IT-Abteilung des endgültigen Betreibers des Informationssystems zusammengearbeitet werden muss.

Die Ergebnisse des Literaturreviews und der Interviews ergeben auch ein einheitliches Bild in Fragen bzgl. des Datenschutzes. Der Schutz der Privatsphäre ist eine prozessübergreifende Herausforderung, betrifft aber insbesondere die Phase Business Understanding, in der entsprechende Regelungen und Fragestellungen mitgedacht werden müssen (Aufgabe Assess Situation). In der Phase Business Understanding müssen rechtliche Rahmenbedingungen erhoben und relevante Fragestellungen für die nächsten Phasen vorbereitet werden. Der Interviewpartner im Bereich Scientific Research hob zusätzlich die Einhaltung ethischer Grundsätze als Herausforderung hervor.

Eine weitere Herausforderung ist die Planung des Projektes (Aufgabe Produce Project Plan). Datenverständnis zu erlagen und die Daten für die Analyse vorzubereiten benötigt einen Großteil der Zeit im Projekt. Im CRISP-DM entspricht dieses Datenverständnis der Phase Data Understanding, für welche im Projektplan entsprechend Zeit eingeplant werden sollte. Diese Einschätzung konnte sowohl bei der Literaturrecherche wie auch in den Interviews gewonnen werden. In den Interviews wurde hervorgehoben, dass auch das Verständnis der Technologie Zeit benötigt, sowohl bei den ExpertInnen, als auch beim KundInnen. Die Schätzung und Planung des Aufwandes ist Voraussetzung für die Schätzung der Kosten. Diese stellen weite Herausforderungen dar, welche in der

Literatur, in den Studien und in den Interviews gleichermaßen genannt werden.

5.2 Data Understanding, Data Preparation, Modeling

Herausforderung	**L**	**I**	**F**
Bestehende Speichertechnologien und Infrastrukturen sind zur Analyse nicht geeignet	Ja	Ja	Ja
Bestehende Analysesoftware zur Analyse nicht geeignet	Ja	Ja	Ja
Datenzugriff und Datenübertragung	Ja	Ja	Nein
Verstehen der Daten	Ja	Ja	Nein
Integration und Vorverarbeitung der Daten	Ja	Ja	Nein, nur Dauer Ladeprozess
Modellierungstechniken (Algorithmen) anpassen und nutzen	Ja	Ja	Ja

Tab. 2: Zentrale Herausforderungen der Phasen Data Understanding, Data Preparation und Modeling

Die Phasen Data Understanding, Data Preparation und Modeling beinhalten laut Franks einen Großteil des Aufwandes [Fr12] in einem Big Data Projekt und sind von übergreifenden technischen Herausforderungen beeinflusst. Zum einen sind die herkömmlichen ggf. bestehenden Technologien wie relationale DBMS, NAS Speichertechnologien oder herkömmliche Analyseverfahren wie OLAP oft nicht geeignet, um mit großen, unstrukturierten Datenmengen in vertretbarer Geschwindigkeit umgehen zu können. Zum anderen müssen neue, aktuelle Technologien durch ExpertInnen verstanden aber auch beherrscht werden. Die Erkenntnisse, dass aktuelle Technologien und das Verständnis dieser benötigt werden, konnte im systematischen Literaturreview gewonnen und durch die Experteninterviews belegt werden.

Die wesentlichen Herausforderungen beim Sammeln der Daten (Aufgabe Collect Initial Data) im Prozess der Wissensgewinnung sind der Datenzugriff und die Datenübertragung. Der Datenzugriff muss verfügbar, erlaubt und sicher sein und ebenso den rechtlichen Anforderungen, wie dem Datenschutzrecht entsprechen. Datenübertragungsgegebenheiten müssen geprüft und dokumentiert werden. Durch eine Vielzahl von Datenquellen in einem Big Data Projekt kann das Management der Datenquellen und die Datenübertragung eine komplexe Herausforderung darstellen. Des Weiteren können Abhängigkeiten zu Schnittstellen (APIs) und darauf aufbauenden Werkzeugen vorhanden und der Zugriff mittels dieser ggf. eingeschränkt sein. Erfahrungen und Fachwissen bei der Benutzung von APIs und Werkzeugen wurden insbesondere im Interview im Bereich Scientific Research als Herausforderung hervorgehoben.

Die weiteren Aufgaben der Phase Data Understanding behandeln das Verstehen der Daten (Aufgaben Describe Data, Explore Data) und die Datenqualität (Aufgabe Vertify

Data Quality). Dieses Verstehen (also das anreichern der Daten mit Kontextinformationen) ist mittels herkömmlicher Datenmodelle meist nicht möglich. Zum einen können Unsicherheiten in den Daten nicht abgebildet werden, zum anderen ist es nur schwer möglich unstrukturierte Daten zu modellieren. Aber nicht nur das Verstehen noch nicht modellierter Datenquellen, sondern auch das Verstehen bestehender Datenstrukturen ist eine Herausforderung, da vorhandene Datenmodelle meist unvollständig sind. Um ein Verständnis der Daten, unabhängig von den Modellen zu gewinnen, müssen Anfragen an diese gestellt werden. Aber auch hier stoßen herkömmliche Anfragetechniken, wie einfaches SQL oder zur herkömmlichen explorativen Datenanalyse eingesetzte Techniken wie OLAP, durch große, unstrukturierte Datenmengen an ihre Grenzen. Diese Art von Daten und die zusätzliche Tatsache, dass es sich meist um viele Datenquellen handelt, macht auch die Sicherstellung der Datenqualität zu einer Herausforderung. Für all diese Tätigkeiten müssen u.U. aktuelle Technologien, z.B. für die Erkundung großer verteilter, unstrukturierter Datenmengen oder die Überprüfung der Datenqualität von großen Datenmengen wie Apache Hadoop oder NoSQL Datenbanken eingesetzt werden.

Wurden die Daten gesammelt und verstanden, müssen diese vorverarbeitet und integriert werden (Phase Data Preparation). Diese Integration und Vorverarbeitung ist für die Erstellung der Modelle ein wichtiger Schritt. Daten müssen mit Informationen angereichert werden (z.B. Metadaten erzeugt oder mit Schlagworten versehen werden). Hierbei ist auch die Beachtung des Datenschutzes wichtig, denn durch gezielte Vorverarbeitungsschritte kann man diesem durchaus gerecht werden. Allerdings sind für all diese Tätigkeiten bei großen, unstrukturierten Datenmengen auch aktuelle Technologien nötig. Insbesondere bei der Benutzung verteilter Systeme ist eine intelligente Aufteilung der Daten notwendig, dies hebt der Interviewpartner im Bereich Commerce and Business mehrfach hervor. Die Vorverarbeitung und Integration der Daten wurde in diesem Interview als zentraler Unterschied zu herkömmlichen Datenanalyseprojekten identifiziert. Auch in der Literatur wurde die bedeutende Rolle dieser Phase hervorgehoben, z.B. [Fr12]. Auffällig ist, dass diese Herausforderungen, bis auf die erhöhte Dauer des Ladeprozesses, nicht in den genannten Studien als solche identifiziert wurden [Gl12], [BI12].

Bei der eigentlichen Analyse (Phase Modeling) ist es eine Herausforderung, die richtige Modellierungstechnik zu wählen. Da die Modellierungstechnik die Art der Vorverarbeitung der Daten bestimmt, sind diese Phasen durchaus mehrfach im Wechsel zu durchlaufen. Aus dem Stand der Forschung geht hervor, dass ein wichtiges Designziel in Big Data Projekten die Unterstützung einer Vielzahl von Analysemethoden ist [BH12]. NutzerInnen soll die Möglichkeit haben Tools und Methoden frei zu wählen. Dies stellt im Prozess der Wissensgewinnung eine große, vor allem, technologische Herausforderung dar. Aber auch die Methoden, also auch die Algorithmen an sich, müssen ggf. weiterentwickelt und auf unstrukturierte, unsichere und verteilte Daten angepasst werden. Dabei muss auch Nebenläufigkeit, Skalierbarkeit und ggf. die Berechnung Vorort (bei den Datenquellen) unterstützt und aktuelle Technologien beachtet werden.

5.3 Evaluation und Deployment

Herausforderung	L	I	F
Evaluation von Modellen und Infrastruktur	Ja	Nein	Nein
Bestehende Speichertechnologien und Infrastrukturen sind zum Betrieb nicht geeignet	Ja	Ja	Ja
Bestehende Analysesoftware und Algorithmen sind im Betrieb nicht geeignet	Ja	Ja	Ja
Visualisierung von Big Data	Ja	Nein	Nein

Tab. 3: Zentrale Herausforderungen der Phasen Evaluation und Deployment

In der Phase Evaluation sollen die Modelle auf ihre Korrektheit überprüft werden. Dies ist durch die großen Datenmengen eine Herausforderung. Auch können Langzeitmodelle nur schwer unmittelbar verifiziert werden. Aber auch die eingesetzte Infrastruktur sollte evaluiert werden. Methoden hierfür sind kaum vorhanden. Auffällig ist, dass Herausforderungen betreffend der Evaluation nur in geringer Anzahl und ausschließlich mittels des Literaturreviews identifiziert werden konnten.

Ein anderes Bild ergab sich bei der Analyse der Herausforderungen in der Phase Deployment. Hier wird die Auslieferung des Modelles geplant. Da es sich in der Praxis, insbesondere durch das Interview im Bereich Commerce and Business unterstützt, meist um die Auslieferung des Analysemodelles an die IT- Abteilung des Kunden, d.h. die Überführung der Prozessergebnisse (Analysemodelle, Parameter der Modelle, Verfahrensweisen, Erfahrungen, usw.) in ein produktives Informationssystem handelt, sind hier viele Herausforderungen identifiziert wurden, die zum einen bei der Planung der Auslieferung auftreten, aber insbesondere bei der Auslieferung direkt. Die zentralen Herausforderungen lassen sich grob in die Bereiche: bestehende Speichertechnologie und Netzwerkinfrastruktur, bestehende Analysesoftware und Visualisierung untergliedern.

Die bestehenden Speichertechnologien und die Netzwerkinfrastruktur muss bei der Planung der Auslieferung mitgedacht werden. Dabei sind Herausforderungen, dass die Anforderungen an Geschwindigkeit, Skalierbarkeit und Performance einer Big Data Analyse meist nicht von den bestehenden/herkömmlichen Speichertechnologien und Netzwerkinfrastrukturen unterstützt werden. Meist sind diese nicht für den verteilten Betrieb ausgelegt (bspw. hinsichtlich Fehlertoleranz, Konsistenz, Hochverfügbarkeit oder Nebenläufigkeit) und ebenso anfällig gegenüber inkonsistenten, unvollständigen oder ungenauen Daten. Beim Einsatz verteilter Systeme, z.B. ein Apache Hadoop Cluster, ist meist auch die vorhandene Netzwerkbandbreite ungenügend, wenn diese durch Map-Reduce-Jobs zusätzlich beansprucht wird.

Ein ähnliches Bild ergibt sich bei der Betrachtung herkömmlicher/bestehender Analysesoftware. Auch hier sind skalierbare Techniken, z.B. für Text Mining oder Soziale Netzwerkanalyse, nötig. Diese müssen ebenfalls parallel arbeiten und werden meist durch eine Middelware verwaltet. Eine Erweiterung bestehender Analysesoftware

muss, unter Beachtung der Ergebnisse im Prozess, geplant werden.

Herausforderungen bei der Planung der Visualisierung konnten ausschließlich durch das systematische Literaturreview gefunden werden. Aber auch Begoli und Horey empfehlen, dass die resultierenden Daten verfügbar und verständlich sein müssen [BH12]. Herausforderungen bei der Visualisierung sind die Multidimensionalität der Daten und die Ungeeignetheit von herkömmlichen Tools und Techniken. Eine Quelle empfiehlt, dass Visualisierung völlig neu gedacht werden muss, anstatt sie an bestehende Ansätze anzupassen [PZ14].

6 Fazit

Durch ein systematisches Literaturreview und Experteninterviews wurden zahlreiche Herausforderungen durch Big Data identifiziert. Alle Herausforderungen konnten einer Phase des CRSIP-DM zugeordnet oder als übergreifende Herausforderung kategorisiert werden. Ein großer Teil der gefundenen Herausforderungen findet sich auch in empirischen Untersuchungen des TDWI Germany e.V. „Big Data in deutschen Unternehmen – Relevanz und Reife“ [Gl12] oder der BITKOM „Big Data im Praxiseinsatz“ [BI12] wieder. Auffällig ist, dass Herausforderungen in den Phasen Data Understanding und Data Preparation, also den Phasen die laut Franks einen Großteil des Aufwandes ausmachen [Fr12] in den genannten empirischen Untersuchungen nicht als Herausforderungen von Unternehmen genannt werden. Dies gilt ebenfalls für Herausforderungen der Evaluation und Visualisierung. Dagegen ist die Mischung aus organisatorischen und technischen Herausforderungen sowohl Ergebnis dieser Untersuchung, als auch der genannten Studien.

Durch die Beschreibung der Herausforderungen in den einzelnen Phasen des CRISP-DM wurde deutlich, dass die Planung und der Einsatz aktueller Technologien entscheidend und eine der wesentlichen Herausforderungen im Prozess sind. Dies findet sich ebenso in den Ergebnissen des systematischen Literaturreviews als auch in den Experteninterviews mehrfach wieder. Dies ist nicht verwunderlich, da Datensätze eben dann als Big Data bezeichnet werden, wenn es mit aktuellen Technologien schwierig ist, diese zu erheben, zu pflegen, zu analysieren und zu visualisieren [PZ14], [Ja11]. Begoli und Horey erkannten das Bereitstellen von Lösungen zur Speicherung und Verarbeitung von Daten in allen Prozessschritten als ein zentrales Designziel in Big Data Projekten [BH12]. Ein Interviewpartner hob hervor, dass eine extra Phase zum Technologieverständnis im Projekt durchgeführt werden musste und diese das Projekt verzögerte. Dies lässt den Schluss zu, dass Technologieentscheidungen frühzeitig, d.h. in der Phase Business Understanding, mitgedacht werden müssen, um die Grundlage späterer Phasen zu legen und diese nicht zu verzögern. Diese ist bisher so im Prozessmodell CRISP-DM nicht im Detail vorgesehen, kann aber durchaus in Aufgaben wie Assess Situation oder Produce Project Plan hineininterpretiert werden. Weiterhin konnte gezeigt werden, dass viele Herausforderungen bei der Auslieferung an sich

auftreten. Hierfür ist in CRISP-DM nur die Planung vorgesehen und es bleibt fraglich, ob dies die Problematik ausreichend abdeckt. Es wurde deutlich, dass der Einsatz eines Prozessmodells zur Wissensgewinnung aus Big Data möglich und sinnvoll ist. Allerdings sollten die aufgezeigten Herausforderungen der einzelnen Phasen beachtet und für das individuelle Projekt eingeplant werden. Dabei sollte, neben der Planung und dem durchgehenden Einsatz von aktuellen Technologien, ein besonderer Fokus auf den Phasen Data Understanding und Data Preparation liegen.

Literaturverzeichnis

[Be10] Berthold, M. R. et al.: Guide to intelligent data analysis. How to intelligently make sense of real data. Springer, London, 2010.

[BH12] Begoli, E.; Horey, J.: Design Principles for Effective Knowledge Discovery from Big Data: 2012 Joint Working IEEE/IFIP Conference on Software Architecture (WICSA) & European Conference on Software Architecture (ECSA), 2012; S. 215–218.

[BI12] BITKOM: Big Data im Praxiseinsatz - Szenarien, Beispiele, Effekte. http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online%281%29.pdf, 09.02.2015.

[BI14] BITKOM: Leitfaden Big-Data-Technologien - Wissen für Entscheider. Leitfaden. http://www.bitkom.org/files/documents/BITKOM_Leitfaden_Big-Data-Technologien-Wissen_fuer_Entscheider_Febr_2014.pdf, 18.04.2015.

[FPS96] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In (Fayyad, U. M. et al. Hrsg.): Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996; S. 1–34.

[Fr12] Franks, B.: Taming the big data tidal wave. Finding opportunities in huge data streams with advanced analytics. John Wiley & Sons Inc, Hoboken, New Jersey, 2012.

[Gl12] Gluchowski, P.: Big Data in deutschen Unternehmen - Relevanz und Reife. www.tdwi.eu/wissen/studien, 01.05.2015.

[Gl14] Gluchowski, P.: Empirische Ergebnisse zu Big Data. In HMD Praxis der Wirtschaftsinformatik, 2014, 51; S. 401–411.

[IB10] IBM: CRISP-DM 1.0 Step-by-step data mining guide. http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=YTW03084USEN, 10.12.2014.

[ID14] IDC: Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm, 18.04.2015.

[Ja11] James Manyika; Michael Chui; Brad Brown; Jacques Bughin; Richard Dobbs; Charles Roxburgh; Angela Hung Byers: Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/~/media/McKinsey/dotcom/

Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx, 18.04.2015.

[Ki04] Kitchenham, B.: Procedures for performing systematic reviews. Keele University, Keele, 2004.

[Ki14] Kitchin, R.: The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE Publications, London, 2014.

[La01] Laney, D.: 3D Data Management Controlling Data Volume, Velocity, and Variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, 13.10.2014.

[Ma10] Mayring, P.: Qualitative Inhaltsanalyse. Grundlagen und Techniken. Beltz, Weinheim, 2010.

[PZ14] Philip Chen, C. L.; Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. In Information Sciences, 2014, 275; S. 314–347.

[Ra13] Rajpurohit, A.: Big data for business managers — Bridging the gap between potential and value: 2013 IEEE International Conference on Big Data, 2013; S. 29–31.